植 物 分 类 学 报 27 (3): 232—239 (1989)
Acta Phytotaxonomica Sinica

分支分类的一种计算方法——最大同步法

徐 克 学
(中国科学院植物研究所)

摘要 本文讨论数量分支分类,提出一种构造演化树的实际计算方法,称为最大同步法。桔梗科 6 个种的数据作为例子进行说明,并对这种方法做简单的评论和比较。

关键词 分支分类;数量分类学

引 言

定量的分支分类是数量分类学的一个重要分支。它与表征分类的定量方法相对立,二者分别来自两种不同生物学分类观点,即表征分类和分支分类。在数量分类学中,表征分类发展较早,理论和方法都比较完善;分支分类起步较晚,方法很不完善。由于分支分类建立在生物演化的谱系关系上,体现了生物系统发育真正的进化关系,因而受到许多生物分类学家的重视。特别在生物类群的系统学研究、进化理论的研究和某些遗传学问题的研究,这些研究都离不开分支分类。

分支分类的生物学理论于 50 年代由德国昆虫学家 W. Hennig 提出。他的一本名为《系统发育分类学》(Phylogenetic Systematics) 一书全面地阐述了分支分类观点。60 年代分支分类很快被引入数量分类学。早期从事分支分类数学方法研究的人有 W. H. Wagner, A. W. F. Edwards 和 L. L. Cavalli-Sforza, 随后有 J. S. Farris, J. H. Camin 和 R. R. Sokal。Camin 和 Sokal 提出了分支分类节省原理(即最短进化路径原理),这一原理为寻求分支分类的分支谱系图奠定理论基础。70 年代又有 G. F. Estabrook 和 F. R. McMorris 等,这些学者从事大量的理论研究工作,其中许多属于数学理论与方法方面的探讨。

当前,分支分类的理论工作有了较大的进步,分支分类的数学理论被构筑在图论和抽象代数的基础上,许多数学家关心抽象概念的引进和数学结论的严谨证明。可是在分类的实践中,为分类学家提供实际使用的方法却为数甚少。Sneath 和 Sokal 的数量分类学经典著作《数量分类学——数量分类的原理和实践》,也只介绍了 Wagner 树和单元法等为数不多的几种方法。随着分支分类应用的发展,这些方法远远不能满足各种生物学问题的需要。分类学家需要更合适的分支分类方法,需要利用电子计算机进行谱系分析的新手段。为此目的,作者从现有的分支理论中引出一种较好的方法,供系统学和分类学研究工作使用。

1987.09.21 收稿。

概 念 与 思 路

与表征分类一样,被分类的实体称为分类运算单位(简作分类单位、或 OTU), 根据推断而得到的演化祖先,称为假设分类单位(简作假设单位, 或 HTU)。为了讨论问题方便,引用统一的名词,把分类单位和假设单位都一律称为分支分类运算单位 (简作分支单位,或 CTU)。

分类学家研究的分类问题,最初由分类单位组成被分类群,如果有 t 个分类单位, 配合 n 个性状。 在此要求对性状状态的演化关系都已分析清楚。 性状的编码都取非负整数,并且规定从 0 开始, 依演化的次序从小到大顺序增加。从这个规定去理解,编码为 0 的性状状态在所研究的范围应该是最原始的状态。 t 个分类单位, n 个性状的全部编码值构成原始数值矩阵:

$$\begin{bmatrix} x_{11} & x_{12} \cdots x_{1n} \\ x_{21} & x_{22} \cdots x_{2n} \\ \cdots\cdots \\ x_{t1} & x_{t2} \cdots x_{tn} \end{bmatrix}$$

其中行代表分类单位向量,列代表性状向量。这个矩阵是分支分类运算的出发点。

为了定量地表示演化程度,引进绝对距离系数。如果有两个分支单位,从原始数值矩阵获得它们的向量表示 $x_i = (x_{i1}\ x_{i2} \cdots x_{in})$ 和 $x_j = (x_{j1}\ x_{j2} \cdots x_{jn})$, 演化距离计算如下:

$$d(x_i, x_j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}| \tag{1}$$

如果性状编码都取整数, 演化距离 d 是非负整数。每个单位值称为 1 步, 作为演化距离的基本单位。

在演化过程中,分支单位 x_p 和 x_q 同时从某个分支单位经 x_r 直接演化而来, 称 x_r 为 x_p 和 x_q 的最近共同祖先。根据分支分类的中位值原则, x_r 的性状分量应取 x_p 和 x_q 相应性状分量的最小值,即:

$$x_{rk} = \min(x_{pk}, x_{qk})\ (k = 1, 2, \cdots, n) \tag{2}$$

生物的进化具有树状演化结构, 以图论中的树图表示生物演化关系是很自然的事。我们把代表生物演化关系的图称为演化图。生物演化不可能出现倒退, 因此演化图在演化的路径上没有回路。还需假设任意两个已经分化了的分支单位不可能再融合而产生网状进化,并且被考虑的分类群是单源的,即都由一个共同的祖先演化而来, 最后演化图中的演化过程还必须与性状的进化保持一致。满足以上条件的演化图实际上是一棵有向树图,称为分支树系图, 见图 1。

演化图中的点代表分支单位称为分支点,两相邻接的分支点 x_i 和 x_j 由具方向的线段 $l = (x_i, x_j)$ 连接, 该连线称为分支线。 分支线的两个端点, 演化开始的一端 x_i 称为起点, 另一端点 x_j 称为终点。整个演化图由许多分支点和相应的分支线构成分支线起点与终点间依公式 (1) 确定的演化距离称为分支线的演化长度。演化图所有分支线的演化长度总和称为演化图的演化长度。

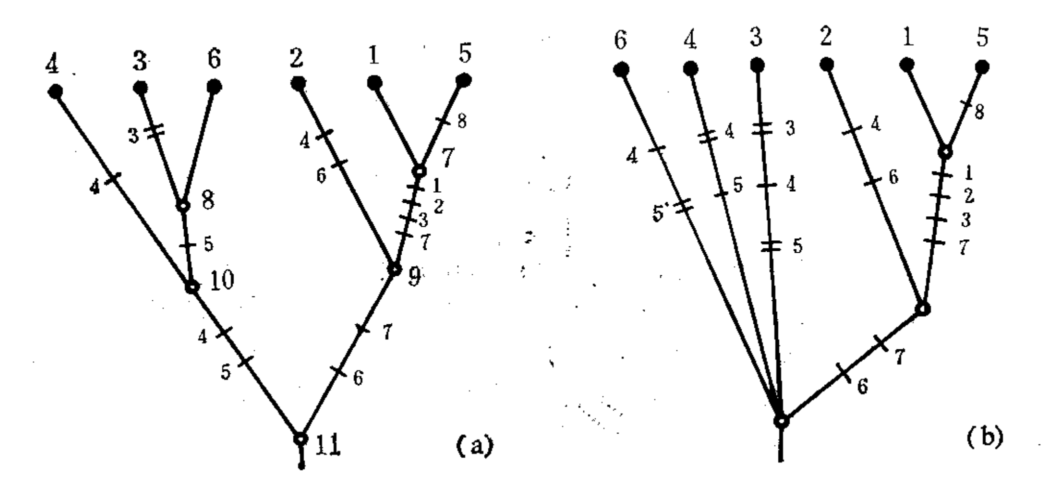


图1 分支树系图：一横截线表示一个演化步，在旁的数字是发生演化的性状。
Fig. 1 Cladogram: A cross-bar indicates an evolutionary step for the character whose number is plased beside it.
(a) 最大同步法，演化长度=15步 $L_{min}=13$ 步
Method of maximal same step length. Evolutionary length = 15 steps. $L_{min}=13$ steps.
(b) 单元法，演化长度=20步 $L_{min}=13$ 步
Monothetic method, Evolutionary length = 20 steps. $L_{min}=13$ steps

代表生物演化关系的演化图，在一切可能构造的图中其演化长度取最小值。这就是所谓最短演化路径原理。为了使演化图的演化长度达到最小值，从公式(1)可知在计算演化长度求和时，应尽量把相同性状状态的进化都表现在同一个分支线上。如果每个性状状态的进化都仅在一个分支线上计入一次，最节省的演化长度必是

$$L_{\min}=\sum_{i=1} m_i$$

其中 m_i 代表第 i 个性状的最大编码值。

为了获得最小演化长度的演化图，引入同步系数的概念，两分支点 $x_i=(x_{i1}\ x_{i2}\cdots x_{in})$ 和 $x_j=(x_{j1}\ x_{j2}\cdots x_{jn})$，值

$$s_{ij}=\sum_{k=1}^{n}\min(x_{ik},\ x_{jk}) \tag{3}$$

称为分支点 x_i 和 x_j 的同步系数。在构造演化图的过程中，将同步系数最大的一对分支点按演化的逆方向优先结合，导出其最近共同祖先。这样的结合将使较多的性状在计算演化长度时仅计入一次，从而达到节省演化长度的要求。这就是最大同步分支分类方法的基本思想。

运算步骤

依照最大同步系数分支单位首先结合的思想，设计分支分类运算。运算在数据矩阵和系数矩阵上以多次循环的过程进行。数据矩阵放置该次循环中被处理的分支单位数据，行代表 CTU，列代表性状。系数矩阵放置被处理 CTUs 之间的同步系数值。行所代

表的 CTU 与数据矩阵完全一致，列所代表的 CTU，排列次序与行完全相同。具体运算步骤规定如下：

1. 按性状状态的进化次序进行性状编码，每个性状的最原始状态取 0 值，其它状态依进化次序从小到大取非负整数，得 t 行 (CTUs) n 列(性状)原始数值矩阵置数据矩阵中。

2. 利用公式 (3) 计算数据矩阵中所有分支单位间的同步系数 $S_{ij}\ (i \neq j)$，置系数矩阵中。上次循环保留的同步系数可以省略计算。

3. 从系数矩阵中找出同步系数最大值。 假如就是 S_{pq}，由此确定把分支单位 x_p 与 x_q 相结合。若有两个以上同步系数达到最大值，可任择一个执行。

4. 根据公式 (2) 求出分支点 x_p 与 x_q 的最近共同祖先 x_r 的性状分量值。从数据矩阵中删除分支单位 x_p 和 x_q 的数据，补充以新的分支单位 x_r，矩阵分支单位(行)数比原来减 1。

5. 在分支树系图上作出从分支点 x_r 到分支点 x_p 和 x_q 的分支关系，并根据公式 (1)

表 1　桔梗科6个种分支分类运算过程

Table. 1　A computing process of cladistic taxonomy on the data of 6 species from campanulaceae

循环次数 Times of cycle	CTUs 编号 No. of CTU	系数矩阵 Coefficient Matrix						性状 Characters	数据矩阵 Data Matrix 1	2	3	4	5	6	7	8
I	1	×							1	1	1	0	0	1	2	0
	2	2	×						0	0	0	1	0	2	1	0
	3	1	1	×					0	0	2	1	2	0	0	0
	4	0	1	2	×				0	0	0	2	1	0	0	0
	5	6	2	1	0	×			1	1	1	0	0	1	2	1
	6	0	1	3	2	0	×		0	0	0	1	2	0	0	0
II	2	×							0	0	0	1	0	2	1	0
	3	1	×						0	0	2	1	2	0	0	0
	4	1	2	×					0	0	0	2	1	0	0	0
	7	2	1	0	×				1	1	1	0	0	1	2	0
	6	1	3	2	0	×			0	0	0	1	2	0	0	0
III	2	×							0	0	0	1	0	2	1	0
	4	1	×						0	0	0	2	1	0	0	0
	7	2	0	×					1	1	1	0	0	1	2	0
	8	1	2	0	×				0	0	0	1	2	0	0	0
IV	9	×							0	0	0	0	0	1	1	0
	4	0	×						0	0	0	2	1	0	0	0
	8	0	2	×					0	0	0	1	2	0	0	0
V	9	×							0	0	0	0	0	1	1	0
	10	0	×						0	0	0	1	1	0	0	0
VI	11								0	0	0	0	0	0	0	0

注：表中原始数据来源见文献 [10]。原数据为举例说明而设，仅取桔梗科少数种；对性状的演化关系也未做认真研究，作为分支演算的例子，姑且认为原编码都符合进化规律。由于以上原因，本例的计算结果不能代表桔梗科的真实情况。请读者注意。

记下该分支线的演化长度和产生演化的性状。

若数据矩阵的分支单位(行)数≥2，则转向步骤2进入下一次循环运算。否则结束运算。

最后检查分支树系图中是否出现演化长度为0的"分支线"。若有，将完全相同的起点与终点重合，取消演化长度为0的"分支线"。

以桔梗科6个种的数据为例，演算过程列表于后。运算结果画出分支树系图(见图中a)

说 明

第 I 次循环运算:

先根据公式(3)计算同步系数。例如计算 S_{12}:

x_1 的性状分量	1	1	1	0	0	1	2	0
x_2 的性状分量	0	0	0	1	0	2	1	0
最小值	0	0	0	0	0	1	1	0

$$S_{12}=\sum_{k=1}^{8}\min(x_{1k},\ x_{2k})=0+0+0+0+0+1+1+0=2。$$

然后根据最大同步系数 $S_{51}=6$ 确定 CTU_5 与 CTU_1 相结合，二者的最近共同祖先是 CTU_7。

在分支树系图上作出分支点 CTU_5、CTU_1 和 CTU_7 和相应的分支线，表示从 CTU_7 演化到 CTU_5 和 CTU_1。

第 II 次循环运算:

根据公式(2)求得 CTU_7 的性状分量值

x_5 的性状分量	1	1	1	0	0	1	2	1
x_1 的性状分量	1	1	1	0	0	1	2	0
x_7 的性状分量(最小值)	1	1	1	0	0	1	2	0

将 CTU_7 的性状分量数据置数据矩阵中并计算 CTU_7 与其它分支单位的同步系数。例如

$$S_{72}=\sum_{k=1}^{8}\min(x_{7k},\ x_{2k})=0+0+0+0+0+1+1+0=2,$$

$$S_{73}=\sum_{k=1}^{8}\min(x_{7k},\ x_{3k})=0+0+1+0+0+0+0+0=1$$

其它同步系数值得自循环 I 中的系数矩阵。

依照本次循环最大同步系数值 $S_{63}=3$，确定 CTU_6 与 CTU_3 相结合，二者的最近共同祖先是 CTU_8。

在分支树系图中补充相应的分支点和分支线，表示从 CTU_8 分别演化到 CTU_6 和 CTU_3 的过程。

第 III 次循环运算:

按公式(2)求 CTU_8 的性状分量值

x_3 的性状分量	0	0	2	1	2	0	0	0
x_6 的性状分量	0	0	0	1	2	0	0	0
x_8 的性状分量(最小值)	0	0	0	1	2	0	0	0

将 CTU_8 的性状分量位置数据矩阵中，并计算 S_{82}，S_{84} 和 S_{87}。例如

$$S_{82}=\sum_{k=1}^{8}\min(x_{8k},\ x_{2k})=0+0+0+1+0+0+0+0=1。$$

其它同步系数来自第 II 次循环。

本次循环的最大同步系数 $S_{77}=S_{84}=2$，任择其中之一。不妨取 S_{72}，由此确定 CTU_7 和 CTU_2 相结合。最近共同祖先是 CTU_9。

在分支树系图中补入从 CTU_9 到 CTU_7 和 CTU_2 的演化过程。

第 IV 次循环运算：

计算 CTU_9 的性状分量值

x_2 的性状分量	0	0	0	1	0	2	1	0
x_7 的性状分量	1	1	1	0	0	1	2	0
x_9 的性状分量(最小值)	0	0	0	0	0	1	1	0

将 CTU_9 的性状分量值记入数据矩阵，并计算同步系数，

$$S_{94}=S_{98}=0,$$

$S_{48}=2$ 得自上次循环。

最大同步系数 $S_{48}=2$，确定 CTU_4 与 CTU_8 相结合，最近共同祖先是 CTU_{10}。

在分支树系图中补入从 CTU_{10} 到 CTU_4 和 CTU_8 的演化过程。

第 V 次循环运算：

计算 CTU_{10} 的性状分量值

x_4 的性状分量	0	0	0	2	1	0	0	0
x_8 的性状分量	0	0	0	1	2	0	0	0
x_{10} 的性状分量(最小值)	0	0	0	1	1	0	0	0

数据矩阵只有 CTU_9 和 CTU_{10}，计算 $S_{910}=0$。将最后两个分支单位结合，产生最近共同祖先 CTU_{11}，即演化图的祖源。

在分支树系图中补入从 CTU_{11} 到 CTU_9 和 CTU_{10} 的演化过程。

第 VI 次循环运算：

从上次循环运算的两个分支单位性状分量得祖源 CTU_{11} 的性状分量值。

数据矩阵保留 CTU 的个数 <2，运算结束。

讨　　论

依照最短演化距离原则，可以确立分支分类的最优分类判别标准。按这个标准，最大同步法可以获得接近最优分类的演化图。桔梗科 6 个种的数据，其演算结果与表征分类的表征树系图比较一致(见文[10])，分支树系图的演化长度 15 步，与 L_{min} 仅差 2 步，说明此结果比较好。与其它方法比较，图中 b 给出了桔梗科 6 个种相同数据使用单元法的运算结果，得到的分支树系图演化长度高达 20 步。一些运算的经验初步证实最大同步法

通常比其它一些分支分类方法能得到较满意的分类结果。

最大同步法演算简单，易于计算机程序化，是其另一个优点。最大同步法没有复杂数学计算，对于一个小规模数据，手工运算也可以完成。利用电子计算机计算，需要编写计算机程序检查全部运算过程，由多次循环运算完成，每次循环所完的运算步骤规律性强，适合编写程序。在设计程序时还会发现，它与表征分类的运算过程有许多相同之处，如果已经有一个表征分类系统聚类运算程序，只需作部分更改就可以得到最大同步法的程序。甚至可以将它与表征分类多种方法统一在一个计算机程序中，这将为分支分类研究工作带来很大的方便。

参 考 文 献

[1] Wagner, W. H., 1961: Problems in the classification of ferns. in Recent advances in botany p. 841—844.

[2] Edwards, A. F. and L. L. Cavalli-Sforza, 1963: Reconstruction of evolutionary trees. in Phenetic and Phylogenetic classification.

[3] Camin, J. H. and R. R. Sokal, 1965: A method for deducing branching sequences in phylogeny. *Evolution*, 19: 311—327.

[4] Farris, J. S., 1970: Methods for computing Wagner trees. *Syst. Zool.*, 19(1): 83—92.

[5] Farris, J. S. et al., 1970: A numerical approach to phylogenetic systematics. *Syst. Zool.* 19(2): 172—189.

[6] Estabrook, G. F., 1968: A general solution in partial orders for the Camin-Sokal model in phylogeny. *J. Theoret. Biol.*, 21: 421—438.

[7] Estabrook, G. F. et al., 1975: An idealized concept of the ture cladistic character. *Math. Biosci.*, 23: 263—272.

[8] McMorris, F. R., 1975: Compatibility criteria for cladistic and qualitative taxonomic characters. in Proceedings 8th conference on numerical taxonomy. p. 399—415.

[9] McMorris, F. R., 1977: On the compatibility of binary qualitative taxonomic characters. *Bull. Math. Biol.*, 39(2): 133—138.

[10] 徐克学，1982：浅谈分类学的数学方法。植物分类学报，20(4)：502—509。

AN ALGORITHM FOR CLADISTIC TAXONOMY ——METHOD OF MAXIMAL SAME STEP LENGTH

Xu Ke-xue

(Institute of Botany, Academia Sinica)

Abstract This paper deals with the numerical cladistic taxonomy. A method for constructing evolutionary tree (method of maximal same step length) is proposed in the applications and practice of cladistic taxonomy. Its algorithm runs as follows:

1) According to the order of evolution, characters are coded with nonnegative integers, producing the original data matrix.

2) Calculate the same step coefficients S_{ij} $(i \neq j)$ by the formula (3) and form the coefficient matrix.

3. Find the maximal value S_{pq} of the same step coefficients in the coefficient matrix.

4) According to the maximal same step length S_{pq}, the most recent common ancestor CTU_r of CTU_p and CTU_q can be determinated by (2).

5) draw the cladistic edges of cladogram representing the evolutionary relationship from

OTU_r to OTU_p and OTU_q.

If the number of CTUs in the data matrix $\leqslant 2$, go to (2); otherwise stop.

An example of 6 species from the family Campanulaceae is given for illustration (See Table 1).

In general case, the evolutionary length of the cladogram obtained by this method is shorter than that by monothetic and other methods. Its algorithm is easily performed and is especially suitable for computerizing.

Key words Cladistic classilfication; Numerical taxonomy